Coreference Resolution for the Biomedical Domain: A Survey

Pengcheng Lu and **Massimo Poesio**
School of Electronical Engineering and Computer Science
Queen Mary University of London, United Kingdom
{pengcheng.lu, m.poesio}@qmul.ac.uk

Abstract

Issues with coreference resolution are one of the most frequently mentioned challenges for information extraction from the biomedical literature. Thus, the biomedical genre has long been the second most researched genre for coreference resolution after the news domain, and the subject of a great deal of research for NLP in general. In recent years this interest has grown enormously leading to the development of a number of substantial datasets, of domain-specific contextual language models, and of several architectures. In this paper we review the state-of-the-art of coreference in the biomedical domain with a particular attention on these most recent developments.

1 Introduction

Coreference resolution is the process of identifying entities in a text and finding all mentions that refer to the same entities. It is a fundamental and challenging NLP task, supporting downstream tasks such as information extraction and question answering.

In the biomedical domain, issues with coreference resolution are one of the most frequently mentioned challenges for information extraction from the biomedical literature (Castano et al. 2002, Miwa et al. 2012). Biomedical coreference resolution has become an essential task to support the discovery of complex information by identifying coreference links in biomedical texts.

In recent years in particular, biomedical coreference resolution has attracted a great deal of attention due both to its great potential for application, and to its theoretical interest e.g., as an application of knowledge embeddings and entity linking. Several biomedical coreference corpora have been made available, especially for protein coreference which is a supporting task for BioNLP 2011 shared task (Nguyen et al., 2011).

Biomedical coreference is quite different from the general domain coreference, such as different markable types. Therefore, domain-specific knowledge is important for bridging the gap. Figure 1 shows two examples of biomedical coreference in the CRAFT-CR dataset (Cohen et al., 2017). In example a), the mention *this protein* refers to the antecedent *the HSPB2 gene product*. To understand the coreference relation, we need the background knowledge that proteins are fundamental encoded by genes. In example b), to correctly resolve the coreference relation between these different biomedical entities, the biomedical-domain knowledge that *cholinergic starburst amacrine cell* is a kind of *interneuron* and belongs to *retinal neuron* are required.

Example a) It must be noted, however, that the alphaA/BKO mice also lack **the HSPB2 gene product** and the contribution of **this protein** to normal lens morphology and functions should not be overlooked.

Example b) However, comprehensive assessment of the Rb/E2f1 double-null rescued retina revealed a differentiation defect in **cholinergic starburst amacrine cells** (SACs). Recent breakthroughs have revealed that **these interneurons** are critical for direction selectivity and developmentally important rhythmic bursts …… **SAC** …… **these retinal neurons** ……

Figure 1: Examples of coreference relations in the CRAFT-CR dataset. The mentions marked red are coreferent.

A large number of coreference models for the biomedical domain have already been proposed, from rule-based models (Castano et al. 2002, Kim and Park 2004, Lin and Liang 2004, Nguyen et al. 2012, Miwa et al. 2012, Kilicoglu and Demner-Fushman 2016, Li et al. 2018), machine learning-

Proceedings of the 4th Workshop on Computational Models of Reference, Anaphora and Coreference (CRAC 2021), pages 12–23
Punta Cana, Dominican Republic, November 10–11, 2021. ©2021 Association for Computational Linguistics

based models (Yang et al. 2004, Torii and Vijay-Shanker 2005, Su et al. 2008, Gasperin 2009, Kim et al. 2011) to recent deep learning-based models (Trieu et al. 2018, Trieu et al. 2019, Li et al. 2021). These models usually integrate biomedical specific information, typically specific rules, pre-trained embeddings and features.

This paper reviews and analyses coreference datasets and models for the biomedical domain, as well as recent biomedical language representation models which can enhance coreference models with domain-specific knowledge. In addition, we conduct experiments to evaluate the ability of these language represetation models for biomedical coreference task.

The structure of this paper is as follows. In Section 2 we briefly provide some background on coreference resolution in the general domain. Section 3 reviews the main datasets used to study biomedical coreference. Overviews of biomedical language representation models and biomedical coreference models are provided in Sections 4 and 5, respectively. Section 6 introduces the methodology of comparing the biomedical language representation models for coreference. Section 7 presents the evaluation results including the performance of previous models and our experiments, and Section 8 concludes.

2 Background

Coreference resolution in the general domain has a long history of being studied from early heuristic-based and rule-based approaches to recent learning-based approaches.

Lee et al. (2017) proposed the first end-to-end neural coreference resolution model which uses LSTM encoder. Based on the end-to-end model, many extensions to the model have been proposed. BERT and SpanBERT were proposed to replace the LSTM encoder and achieved better performance on OntoNotes dataset (Joshi et al. 2019, Joshi et al. 2020). Wu et al. (2020) adapted question-answering framework on coreference resolution, and achieved the state-of-the-art result with 83.1% F1 score on OntoNotes dataset. Ye et al. (2020) proposed a novel language representation model CorefBERT, which can capture the coreferential relations in context.

However, these general coreference systems do not work well in the biomedical domain due to the lack of domain knowledge. For example, the end-to-end model (Lee et al., 2017) only achieved 33.85% and 61.25% F1 scores on CRAFT-CR and BioNLP datasets respectively (Trieu et al., 2018), but achieved 68.8% F1 score on OntoNotes dataset (Hovy et al., 2006), which covers multiple genres, such as newswire, broadcast news and web data.

3 Biomedical Coreference Datasets

Several biomedical datasets with coreference annotations exist, but different document selection criteria, annotation schemes, domains and coreference types were used. The best known include:

MEDSTRACT (Pustejovsky et al., 2002) is a corpus consisting of MEDLINE abstracts with coreference annotation. It is mainly concerned with two forms of anaphora: pronominal and sortal (definite noun phrase) anaphora. This corpus adapted the MUC-7 annotation scheme (Hirschman, 1997); in addition, semantic types from UMLS (Bodenreider, 2004) were also annotated.

FlySlip (Gasperin et al., 2007) contains anaphoric links among noun phrases, including coreferent and associative relations. Different from MEDSTRACT, full-text biomedical articles were annotated in this corpus. FlySlip was annotated according to a domain-specific annotation scheme.

GENIA-MedCo (Su et al., 2008) is a coreferentially annotated version of the GENIA corpus (Kim et al., 2003), which in turn consists of 1999 MEDLINE abstracts. This corpus follows the MUC-7 annotation scheme, but adds more linguistic based relations.

DrugNerAR (Segura-Bedmar et al., 2010) was created to study anaphoric expressions in the task of extracting drug-drug interactions in pharmacological literature. This corpus consists of 49 full-text from the DrugBank database, which contains 4900 drug entries.

BioNLP-ST'11 COREF (Nguyen et al., 2011) was created in support of one of the tasks of the BioNLP 2011 shared task, focusing on finding anaphoric protein references, and based on the observation that one of major difficulties in event extraction is coreference resolution. This corpus was derived from three resources: MedCo coreference annotation (Su et al., 2008), Genia event annotation (Kim et al., 2008), and Genia Treebank (Tateisi et al., 2005).

HANAPIN (Batista-Navarro and Ananiadou, 2011) is comprised of 20 full-text articles from biochemistry literature. In addition to nominal and

Dataset	Document Type	Annotation Scheme	Domain	Coreference type	Semantic type of markables	Relation Type	Publicly Available
MEDSTRACT	100 MEDLINE abstracts	MUC-7	molecular biology	pronoun, sortal	UMLS semantic types	pairwise coreference	Yes
FlySlip	5 full-text articles	domain-specific	fruit fly genomics	noun phrase	genetic entities	coreference chain	Yes
GENIA-MedCo	1999 MEDLINE abstracts	MUC-7	transcription factors in human blood cells	pronoun, noun phrase	GENIA Ontology types	pairwise coreference	Yes
DrugNerAR	49 full articles	MUC-7	drug-drug interactions	pronoun, noun phrase	drugs	pairwise coreference	Yes
BioNLP 11 COREF	1210 MEDLINE abstracts	BioNLP	molecular biology	pronoun, noun phrase	protein names	pairwise coreference	Yes
HANAPIN	20 full-text articles	MedCo	marine natural products chemistry	pronoun, sortal, numerical, abbreviation	chemical compounds, organisms, drug effects, diseases, drug targets	coreference chain	Yes
CRAFT-CR	97 full texts articles	OntoNotes	mouse genomics	pronoun, noun phrase, verb, event, Nominal premodifiers	all types	coreference chain	Yes

Table 1: Comparison of biomedical datasets with coreferent annotations.

pronominal anaphora, this corpus also annotated abbreviation/acronyms and numerical anaphora.

CRAFT-CR (Cohen et al., 2017) consists of 97 full-text biomedical journal articles. Similar to the general domain, this corpus was annotated with coreferent chains in full-text articles, while most other biomedical coreference datasets focuse on annotating the pairwise coreference relation between an anaphor and its antecedent. In addition, all coreference expressions were annotated regardless of semantic type.

These datasets are summarized in Table 1.

4 Biomedical Language Representation Models

The news domain and the biomedical domain are different in a number of respects, such as markable types. Some authors have argued that biomedical domain knowledge is the key to bridging the gap (Choi et al., 2014), and that therefore, incorporating biomedical specific representation is beneficial for resolving coreferring expressions in the biomedical domain. In this section, we will give a brief introduction to biomedical language representation models.

4.1 Pre-training on biomedical corpora

Following the success of large-scale pre-training language models (PLMs) in the general domain, several biomedical-domain PLMs have been developed in recent years by pre-training on large-scale biomedical corpora.

Most biomedical PLMs conduct continual pre-training of the general domain PLMs and still use vocabulary trained on the general domain text. BioBERT (Lee et al., 2020) is the first transformer-based biomedical PLM, pre-trained on PubMed abstracts and PubMed Central full-text articles. ClinicalBERT and Bio_ClinicalBERT (Alsentzer et al., 2019) are pre-trained on MIMIC-III Clinical Notes, whereas BlueBERT (Peng et al., 2019) uses both PubMed and MIMIC-III for pre-training. All these models are pre-trained based on general BERT, except Bio_ClinicalBERT which is initialized from BioBERT.

In addition to initializing from general BERT, some biomedical PLMs are directly pre-trained on biomedical text from scratch and use domain-specific custom vocabulary. SciBERT (Beltagy et al., 2019) is pre-trained on biomedical and computer science papers from scratch and achieved good performance on many scientific NLP tasks. PubMedBERT (Gu et al., 2020) and BioELECTRA (raj Kanakarajan et al., 2021) are both pre-trained on PubMed abstract and PubMed Central full text articles, but the latter adopts ELECTRA architecture (Clark et al., 2019). BioMegatron (Shin et al., 2020) is a large-scale model based on Megatron (Shoeybi et al., 2019) architecture. It also investigated the effect of vocabulary and corpora domain on the performance of biomedical tasks.

4.2 Integrating biomedical knowledge bases

Although the biomedical PLMs, such as BioBERT, have achieved good performance on many biomedical tasks, however, these models can be further enhanced by integrating biomedical knowledge bases, such as UMLS (Bodenreider, 2004).

Several models enhance biomedical PLMs by integrating synonym knowledge from UMLS. Each mention in the biomedical text can be linked to a Concept Unique Identifier (CUI) in UMLS, and each CUI has a synonym set. SAPBERT (Liu et al., 2021), UMLSBERT (Michalopoulos et al., 2021) and BIOSYN (Sung et al., 2020) further pre-trained PubMedBERT, Bio_Clinical BERT and BioBERT on UMLS synonyms, using multi-similarity loss, multi-label loss and synonym marginalization algorithm respectively.

In addition to synonym knowledge, Clinical KB-BERT (Hao et al., 2020) injects UMLS relation knowledge into BioBERT. Whereas CODER (Yuan et al., 2020) learns both synonym and relation knowledge based on PubMedBERT or mBERT (Devlin et al., 2019) via contrastive learning. Also, some research focus on fusing the UMLS entity embeddings with contextual embeddings to improve biomedical PLMs (He et al., 2020; Fei et al., 2021; Yuan et al., 2021).

This paper selected some of the models above to evaluate the ability of biomedical-specific representation for biomedical coreference task, detailed in Section 6.

5 Coreference Models for the Biomedical Domain

5.1 Rule-based models

Early approaches to biomedical coreference resolution are primarily rule-based. These models rely on syntactic parsers to extract hand-crafted features and rules.

Nguyen et al. (2012) implemented a protein coreference system that makes use of syntactic information from the parser output, and protein-indicated information. The results showed that domain-specific semantic information is important for coreference resolution. Miwa et al. (2012) developed a rule-based coreference system, as a part of the EventMine event extraction system. A set of rules was developed based on syntactic trees and predicate-argument structures. The system achieved 55.9% F1 score on BioNLP 2011 protein coreference task. Kilicoglu and Demner-Fushman (2016) developed a new corpus of structured drug labels and proposed a general framework based on a smorgasbord architecture for fine-grained biomedical coreference resolution. The framework adopted different strategies for each coreference type and mention type, and combined them to reach desired performance, like selecting dishes from a smorgasbord. Li et al. (2018) presented two methods for bio-entity coreference resolution: a rule-based method and a recurrent neural network (RNN) model. The rule-based model created a set of syntactic rules or semantic constraints for coreference and achieved a state-of-the-art performance with 62.0% F1 score on BioNLP 2011 protein coreference task.

These rule-based models mostly designed rules for specific type of coreference relation and even specific corpus, which limits the scope of the resolution.

5.2 Machine learning-based models

In the early years, due to the lack of publicly available annotated corpora, researchers have to annotate their own corpora for developing machine learning approaches (Yang et al., 2004, Torii and Vijay-Shanker, 2005, Su et al., 2008, Gasperin, 2009).

After the BioNLP 2011 protein coreference dataset was made publicly available, several machine learning-based models were developed for this task. Kim et al. (2011) adapted a general coreference system Reconcile (Stoyanov et al., 2010) for the biomedical domain by modifying several components to biomedical texts. It trained two separate classifiers for detecting anaphora and antecedent mentions.

In addition to using machine learning-based methods only, several models adopted hydrid approach, i.e., combining both machine learning-based and rule-based methods. D'Souza and Ng (2012) proposed a hybrid approach that used a classifier with syntactic path-based features. It investigated five different learning-based methods, and a rule-based approach for anaphora resolution. This model achieved a superior performance than previous either rule-based or learning-based models on BioNLP 2011 protein coreference task. Li et al. (2014) later also used a hybrid approach, adopting the rule-based method or the machine learning method for three types of anaphora. As the method of D'Souza and Ng (2012), they also used different rules for different types of anaphora. The system achieved better performance with 68.6% F1 score than previous methods on BioNLP 2011 protein coreference development data.

	BioNLP	**CRAFT**
Training set (docs)	800	60
Development set (docs)	150	7
Test set (docs)	260	30
Avg. sent. per doc	9.2	312.4
Avg. words per doc	258.0	8181.0

Table 2: Statistics of BioNLP and CRAFT.

5.3 Deep learning-based models

In recent years, much effort has been made on using deep learning methods for biomedical coreference.

Trieu et al. (2018) applied general domain end-to-end neural coreference resolution system (Lee et al., 2017) to biomedical text, integrating the domain specific features to enhance the system. The model was evaluated on BioNLP 2011 protein coreference dataset and CRAFT-CR dataset. The results indicated that in-domain embeddings and domain-specific features helped improve the performance. Then, Trieu et al. (2019) proposed a system to address the challenge of coreference resolution in the full-text articles in the CRAFT-CR dataset. The model also applied end-to-end system (Lee et al., 2017), but enhanced the system by utilizing a syntax-based mention filtering method and replacing LSTM with BERT. This model achieved better performance on the CRAFT-CR dataset.

Different from the models above, Li et al. (2021) integrated external knowledge to enhance the neural coreference system for biomedical texts. A knowledge attention module was developed to select the most related and helpful knowledge triplets. This model achieved the state-of-the-art perfor-

rule-based	Classification machine learning	deep learning	Model	dev R	dev P	dev F1	test R	test P	test F1
	✓		Reconcile (Kim et al., 2011)	26.7	74.0	39.3	22.2	73.3	34.1
✓			(Nguyen et al., 2012)	57.8	67.8	62.4	52.5	50.2	51.3
✓			EventMine (Miwa et al., 2012)	53.5	69.8	60.5	50.4	62.7	55.9
✓	✓		(D'Souza and Ng, 2012)	59.9	77.1	67.4	55.6	67.2	60.9
✓	✓		(Li et al., 2014)	69.8	67.5	68.6	-	-	-
✓			Simple system (Choi et al., 2016)	64.4	63.4	63.9	50.0	46.3	48.1
✓			(Kilicoglu and Demner-Fushman, 2016)	63.2	72.4	67.5	-	-	-
✓			(Li et al., 2018)-rule	68.8	76.0	**72.2**	60.2	63.8	62.0
		✓	(Li et al., 2018)-neural	60.4	61.9	61.2	54.9	58.0	56.4
		✓	E2E_MetaMap (Trieu et al., 2018)	56.7	71.7	63.1	47.5	55.6	51.2
		✓	KB-attention (Li et al., 2021)	63.4	68.1	65.6	69.4	69.6	**69.5**

Table 3: Performance of biomedical coreference models on BioNLP 2011 protein correference development and test sets.

mance on the BioNLP 2011 protein coreference dataset and CRAFT-CR dataset.

6 Comparing the Biomedical Language Representation Models for Coreference

In Section 4, we introduced a series of biomedical language representation models. To investigate the ability of these models for biomedical coreference task, we conduct experiments to evaluate these models on CRAFT-CR dataset.

6.1 Baseline model

We employ the higher-order coreference model (Lee et al., 2018) as the baseline model, but use different pre-trained language models with BERT architecture to replace LSTM encoder.

The goal is to learn a distribution $P(y_i)$ over possible antecedents $Y(i)$ for each span i:

$$P(y_i) = \frac{e^{s(i,y_i)}}{\sum_{y' \in Y(i)} e^{s(i,y')}} \quad (1)$$

where $s(i,j)$ is a pairwise score for a coreference link between span i and span j. The pairwise score is computed by the mention score of i, the mention score of j, and two kinds of joint compatibility scores of i and j:

$$s(i,j) = s_m(i) + s_m(j) + s_c(i,j) + s_a(i,j) \quad (2)$$

The mention score and joint compatibility scores are computed using span representation g_i and g_j from bidirectional LSTMs:

$$s_m(i) = \mathit{FFNN}_m(g_i) \quad (3)$$

$$s_c(i,j) = g_i^T W_c \, g_j \quad (4)$$

$$s_a(i,j) = \mathit{FFNN}_a([g_i, g_j, g_i \circ g_j, \phi(i,j)]) \quad (5)$$

where *FFNN(·)* represents a feed-forward neural network, W_c is a learned weight matrix, $\circ$ denotes element-wise product, and $\phi(i,j)$ represents speaker and metadata features.

6.2 Applying pre-trained language models

We apply two types of PLM to replace LSTM encoder respectively:

Biomedical PLMs: to enhance the baseline model with biomedical domain knowledge, several biomedical PLMs are selected, including models pre-training on biomedical corpora or integrating biomedical knowledge bases.

SpanBERT: since SpanBERT (Joshi et al., 2020) is a state-of-the-art coreference resolution model in the general domain, we also evaluate SpanBERT and general BERT (Joshi et al., 2019) on biomedical coreference.

Since CRAFT-CR is a more challenging biomedical coreference dataset consisting of full-text articles, we choose CRAFT-CR to fine-tune and evaluate these models. The details are introduced in Section 7.2.

7 Results

In this section, we first present the performance achieved by previous biomedical coreference models decribed in Section 5. Then we describe our experiment and report the results.

7.1 Results by datasets

Recent biomedical coreference models are mostly evaluated on BioNLP 2011 protein coreference

Model	B^3	BLANC	CEAFE	CEAFM	LEA	MUC	Avg.
E2E_MetaMap (Trieu et al., 2018)	36.4	46.5	33.1	41.0	32.4	51.8	40.2
BERT_filter (Trieu et al., 2019)	44.0	48.9	39.8	49.0	40.0	57.0	46.4
KB-attention (Li et al., 2021)	54.9	63.1	48.6	59.4	51.3	64.5	**57.0**

Table 4: F1 scores of biomedical coreference models on CRAFT-CR test set.

Model	B^3	BLANC	CEAFE	CEAFM	LEA	MUC	Avg.
BioBERT	41.67	42.39	32.44	45.15	39.00	53.66	42.38
SciBERT	25.66	28.30	16.76	30.34	22.67	40.70	27.41
Bio_ClinicalBERT	38.19	36.91	30.11	41.56	35.57	48.22	38.43
PubMedBERT	34.96	33.14	25.49	38.49	32.32	47.02	35.24
UMLSBERT	27.53	26.95	19.95	31.40	24.95	39.80	28.43
Clinical KB-BERT	44.56	44.99	37.25	48.29	41.67	55.17	45.32
BERT_base	32.96	31.36	22.58	36.25	30.73	44.34	33.04
SpanBERT_base	47.05	46.30	39.90	51.27	44.36	57.69	**47.76**

Table 5: F1 scores of different PLMs combined with c2f-coref model on CRAFT-CR test set.

dataset[1] and CRAFT-CR dataset[2], of which the statistics are shown in Table 2. The performance on the two datasets are summarized and analysed respectively as follows.

Table 3 shows the performance of different biomedical coreference models on BioNLP 2011 protein coreference development and test sets. These models are evaluated using the scorer provided by the BioNLP shared task organisers. As shown in Table 3, KB-attention (Li et al., 2021) achieved the best performance of 69.5% F1 score on the test set of BioNLP. This indicates that integrating external biomedical knowledge base can further enhance the coreference models for the biomedical domain. In addition, compared with deep learning-based models, some rule-based (Kilicoglu and Demner-Fushman 2016, Li et al. 2018) or hybrid models (D'Souza and Ng 2012, Li et al. 2014) still achieved favorable performance.

Table 4 shows the F1 scores of different biomedical coreference models on CRAFT-CR test set. We can see that the best performance is also achieved by KB-attention (Li et al., 2021), showing the advantage of fine-grained knowledge base integration. However, the results on CRAFT-CR are overall lower than those on BioNLP. The possible reason is that CRAFT-CR consists of full-text articles, hence the length of documents in CRAFT-CR is much greater. This makes CRAFT-CR more challenging than BioNLP dataset which comprises abstracts only.

7.2 Experiments

7.2.1 Experimental setup

We conduct experiment using following models:

- **biomedical PLMs+c2f-coref**: we refer to the higher-order coreference model (Lee et al., 2018) as *c2f-coref*. We build the c2f-coref system on top of different biomedical PLMs respectively, including BioBERT, SciBERT, Bio_ClinicalBERT, PubMedBERT, UMLSBERT, and Clinical KB-BERT. Among these models, UMLSBERT and Clinical KB-BERT integrate external biomedical knowledge base, i.e., UMLS, while other models are pre-trained on large-scale biomedical datasets.

- **BERT_base+c2f-coref** (Joshi et al., 2019): the c2f-coref system on top of BERT representation.

- **SpanBERT_base+c2f-coref** (Joshi et al., 2020): the c2f-coref system on top of SpanBERT_base, which pre-trained span representations to better represent and predict spans of text.

We run these models on the CRAFT-CR dataset of latest released version 4.0.1[3]. CRAFT-CR con-

[1] http://2011.bionlp-st.org/home/protein-gene-coreference-task

[2] https://github.com/UCDenver-ccp/craft-shared-tasks

[3] https://github.com/UCDenver-ccp/CRAFT/releases/tag/v4.0.1

Model	Evaluation script	Programming language	Time cost	MUC	B^3	CEAFE	Avg.
SpanBERT_base +c2f-coref	CoNLL scorer 9.0	Perl	about 1.5h	48.67	8.61	18.90	25.39
	CoVal script	Python	about 30s	55.99	43.97	40.01	46.66
	CRAFT evaluation script	Clojure	about 20m	57.69	47.05	39.90	48.21

Table 6: F1 scores of SpanBERT_base+c2f-coref on CRAFT-CR test set using different evaluation scripts.

sists of 97 full-text journal articles from PMC. As shown in Table 2, 60 documents are used for fine-tuning these models.

These models are fine-tuned using learning rate of 1×10^{-5} for PLMs parameters and 2×10^{-4} for task parameters with Adam optimizer, a dropout of 0.3, and `max_training_len` of 384 for SpanBERT_base and 128 for other PLMs respectively. For SciBERT and PubMedBERT, we use the specific domain vocabulary, while general BERT vocabulary is used for other models.

For evaluation, we calculate F1 scores on six common metrics including B^3, BLANC, CEAFE, CEAFM, LEA and MUC using the official evaluation script[4] provided by the CRAFT shared task organizers, which is also used by previous models (Trieu et al. 2018, Trieu et al. 2019, Li et al. 2021).

7.2.2 Results

The F1 scores of different PLMs combined with c2f-coref model on the CRAFT-CR test set are shown in Table 5. We can see that SpanBERT_base achieved the best performance of 47.76% F1 score, even without biomedical domain pre-training. This proves the powerful ability of SpanBERT on coreference resolution task.

In addition, biomedical PLMs outperform BERT_base on CRAFT-CR, except SciBERT (Beltagy et al., 2019) and UMLSBERT (Michalopoulos et al., 2021), which shows that biomedical domain knowledge can generally benefit coreference models for the biomedical domain. Moreover, Clinical KB-BERT (Hao et al., 2020), which is initialized from BioBERT (Lee et al., 2020), achieved better performance than other biomedical PLMs, indicating that biomedical PLMs can be further enhanced by integrating external biomedical knowledge bases. However, SciBERT performs worse than BERT_base on the CRAFT-CR dataset, although pre-trained on scientific texts and achieved better performance than BERT_base on some other scientific NLP tasks such as NER, as reported in Beltagy et al. (2019). One possible reason is that the pre-training corpora of SciBERT contain a number of computer science articles, which is unlikely to be beneficial for biomedical tasks.

Among these biomedical PLMs, SciBERT (Beltagy et al., 2019) and PubMedBERT (Gu et al., 2020) are pre-trained on domain-specific text from scratch, while others conduct continual pre-training based on the general domain. Although Gu et al. (2020) shows that domain-specific pre-training from scratch outperforms continual pre-training from general-domain language models, the results of our experiment are the opposite. Presumably the reason is that CRAFT-CR annotated all semantic type markables and covered a wider range of coreferences, so pre-training on the general domain is also beneficial.

7.2.3 Results using different evaluation scripts

Apart from the official evaluation script provided by CRAFT shared task organizers, we also used two other evaluation scripts, i.e., CoNLL scorer 9.0 and the CoVal script, to evaluate these models for comparing the differences between these evaluation scripts on the CRAFT-CR dataset. CoNLL scorer 9.0[5] is a modified version of the original reference coreference scorer (Pradhan et al., 2014) used for CoNLL-2011/2012 shared tasks. It added an optional partial mention matching scheme and handling for discontinuous mentions, i.e., mentions composed of non-contiguous tokens. The CoVal script[6] is a python coreference scorer for both CoNLL and ARRAU datasets (Uryupina et al., 2020).

The results of three different evaluation scripts for SpanBERT_base+c2f-coref model on CRAFT-CR test set are shown in Table 6. The F1 scores of MUC, B^3 and CEAFE metrics as well as the averaged value are provided. As shown in Table 6, a strange phenomenon is that the results of CoNLL

[4] https://github.com/UCDenver-ccp/craft-shared-tasks

[5] https://github.com/bill-baumgartner/reference-coreference-scorers

[6] https://github.com/ns-moosavi/coval

scorer 9.0 are much lower than those of the other two evaluation scripts, especially on the B^3 and CEAFE metrics. The reason of that is not clear and needs further analysis. Whereas, the results of the CoVal script and CRAFT official evaluation script are close, although the scores of the latter are a little higher.

In addition to the F1 scores, the time cost of the three evaluation scripts are quiet different. The CoNLL scorer 9.0 took about one and a half hours, while the CoVal script only needed about 30 seconds for evaluation.

8 Conclusion

In this paper, we review and analyse the progress of biomedical coreference datasets, biomedical language representation models and coreference models for the biomedical domain. Biomedical coreference is an essential but challenging task. Some efforts have been made in this field, but there is still a much room for improvement. The experiments which we conducted indicate biomedical domain knowledge from either pre-training on biomedical texts or integrating biomedical knowledge bases can enhance coreference models for the biomedical domain.

Acknowledgements

This research was supported in part by the China Scholarship Council, and the DALI project, ERC Grant 695662.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78.

Riza Theresa Batista-Navarro and Sophia Ananiadou. 2011. Building a coreference-annotated corpus from the domain of biochemistry. In *Proceedings of BioNLP 2011 Workshop*, pages 83–91.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

José Castano, Jason Zhang, and James Pustejovsky. 2002. Anaphora resolution in biomedical literature.

Miji Choi, Karin Verspoor, and Justin Zobel. 2014. Evaluation of coreference resolution for biomedical text. In *MedIR@ SIGIR*.

Miji Choi, Justin Zobel, and Karin Verspoor. 2016. A categorical analysis of coreference resolution errors in biomedical texts. *Journal of biomedical informatics*, 60:309–318.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

K Bretonnel Cohen, Arrick Lanfranchi, Miji Joo-young Choi, Michael Bada, William A Baumgartner, Natalya Panteleyeva, Karin Verspoor, Martha Palmer, and Lawrence E Hunter. 2017. Coreference annotation and resolution in the colorado richly annotated full text (craft) corpus of biomedical journal articles. *BMC bioinformatics*, 18(1):1–14.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Jennifer D'Souza and Vincent Ng. 2012. Anaphora resolution in biomedical literature: a hybrid approach. In *Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, pages 113–122.

Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. 2021. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3):bbaa110.

Caroline Gasperin, Nikiforos Karamanis, and Ruth Seal. 2007. Annotation of anaphoric relations in biomedical full-text articles using a domain-relevant scheme. In *Proceedings of DAARC*, volume 2007. Citeseer.

Caroline V Gasperin. 2009. Statistical anaphora resolution in biomedical texts. Technical report, University of Cambridge, Computer Laboratory.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann,

Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*.

Boran Hao, Henghui Zhu, and Ioannis Paschalidis. 2020. Enhancing clinical bert embedding using a biomedical knowledge base. In *Proceedings of the 28th international conference on computational linguistics*, pages 657–661.

Bin He, Di Zhou, Jinghui Xiao, Xin Jiang, Qun Liu, Nicholas Jing Yuan, and Tong Xu. 2020. Integrating graph contextualized knowledge into pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2281–2290.

Lynette Hirschman. 1997. Muc-7 coreference task definition, version 3.0. *Proceedings of MUC-7, 1997*.

Eduard Hovy, Mitch Marcus, Martha Palmer, Lance Ramshaw, and Ralph Weischedel. 2006. Ontonotes: the 90% solution. In *Proceedings of the human language technology conference of the NAACL, Companion Volume: Short Papers*, pages 57–60.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel S Weld. 2019. Bert for coreference resolution: Baselines and analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5803–5808.

Halil Kilicoglu and Dina Demner-Fushman. 2016. Bioscores: A smorgasbord architecture for coreference resolution in biomedical text. *PloS one*, 11(3):e0148538.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Jin-Dong Kim, Tomoko Ohta, and Jun'ichi Tsujii. 2008. Corpus annotation for mining biomedical events from literature. *BMC bioinformatics*, 9(1):1–25.

Jung-Jae Kim and Jong C Park. 2004. Bioar: Anaphora resolution for relating protein names to proteome database entries. In *Proceedings of the Conference on Reference Resolution and Its Applications*, pages 79–86.

Youngjun Kim, Ellen Riloff, and Nathan Gilbert. 2011. The taming of reconcile as a biomedical coreference resolver. In *Proceedings of BioNLP Shared Task 2011 Workshop*, pages 89–93.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692.

Chen Li, Zhiqiang Rao, Qinghua Zheng, and Xiangrong Zhang. 2018. A set of domain rules and a deep network for protein coreference resolution. *Database*, 2018.

Lishuang Li, Liuke Jin, Zhenchao Jiang, Jing Zhang, and Degen Huang. 2014. Coreference resolution in biomedical texts. In *2014 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 12–14. IEEE.

Y Li, X Ma, X Zhou, P Cheng, K He, and C Li. 2021. Knowledge enhanced lstm for coreference resolution on biomedical texts. *Bioinformatics (Oxford, England)*.

Yu-Hsiang Lin and Tyne Liang. 2004. Pronominal and sortal anaphora resolution for biomedical literature. In *Proceedings of the 16th Conference on Computational Linguistics and Speech Processing*, pages 101–109.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4228–4238.

George Michalopoulos, Yuanxin Wang, Hussam Kaka, Helen Chen, and Alexander Wong. 2021. Umlsbert: Clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1744–1753.

Makoto Miwa, Paul Thompson, and Sophia Ananiadou. 2012. Boosting automatic event extraction from the literature using domain adaptation and coreference resolution. *Bioinformatics*, 28(13):1759–1765.

Ngan Nguyen, Jin-Dong Kim, Makoto Miwa, Takuya Matsuzaki, and Junichi Tsujii. 2012. Improving protein coreference resolution by simple semantic classification. *BMC bioinformatics*, 13(1):1–12.

Ngan Nguyen, Jin-Dong Kim, and Jun'ichi Tsujii. 2011. Overview of the protein coreference task in bionlp shared task 2011. In *Proceedings of the BioNLP Shared Task 2011 Workshop*, pages 74–82. Citeseer.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65.

Sameer Pradhan, Xiaoqiang Luo, Marta Recasens, Eduard Hovy, Vincent Ng, and Michael Strube. 2014. Scoring coreference partitions of predicted mentions: A reference implementation. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2014, page 30. NIH Public Access.

James Pustejovsky, José Castano, Roser Sauri, Jason Zhang, and Wei Luo. 2002. Medstract: creating large-scale information servers from biomedical texts. In *Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain*, pages 85–92.

Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. Bioelectra: Pretrained biomedical text encoder using discriminators. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 143–154.

Isabel Segura-Bedmar, Mario Crespo, César de Pablo-Sánchez, and Paloma Martínez. 2010. Resolving anaphoras for the extraction of drug-drug interactions in pharmacological documents. In *BMC bioinformatics*, volume 11, pages 1–9. BioMed Central.

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020. Bio-megatron: Larger biomedical domain language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4700–4706.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. 2019. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*.

Veselin Stoyanov, Claire Cardie, Nathan Gilbert, Ellen Riloff, David Buttler, and David Hysom. 2010. Coreference resolution with reconcile. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 156–161.

Jian Su, Xiaofeng Yang, Huaqing Hong, Yuka Tateisi, and Jun'ichi Tsujii. 2008. Coreference resolution in biomedical texts: a machine learning approach. In *Dagstuhl Seminar Proceedings*. Schloss Dagstuhl-Leibniz-Zentrum für Informatik.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3641–3650.

Yuka Tateisi, Akane Yakushiji, Tomoko Ohta, and Jun'ichi Tsujii. 2005. Syntax annotation for the genia corpus. In *Companion Volume to the Proceedings of Conference including Posters/Demos and tutorial abstracts*.

Manabu Torii and K Vijay-Shanker. 2005. Anaphora resolution of demonstrative noun phrases in medline abstracts. In *Proceedings of*, pages 332–339.

Hai-Long Trieu, Anh-Khoa Duong Nguyen, Nhung Nguyen, Makoto Miwa, Hiroya Takamura, and Sophia Ananiadou. 2019. Coreference resolution in full text articles with bert and syntax-based mention filtering. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 196–205.

Hai Long Trieu, Nhung TH Nguyen, Makoto Miwa, and Sophia Ananiadou. 2018. Investigating domain-specific information for neural coreference resolution on biomedical texts. In *Proceedings of the BioNLP 2018 workshop*, pages 183–188.

Olga Uryupina, Ron Artstein, Antonella Bristot, Federica Cavicchio, Francesca Delogu, Kepa J Rodriguez, and Massimo Poesio. 2020. Annotating a broad range of anaphoric phenomena, in a variety of genres: the arrau corpus. *Natural Language Engineering*, 26(1):95–128.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. Corefqa: Coreference resolution as query-based span prediction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6953–6963.

Xiaofeng Yang, Jian Su, Guodong Zhou, and Chew Lim Tan. 2004. An np-cluster based approach to coreference resolution. In *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pages 226–232.

Deming Ye, Yankai Lin, Jiaju Du, Zhenghao Liu, Peng Li, Maosong Sun, and Zhiyuan Liu. 2020. Coreferential reasoning learning for language representation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7170–7186.

Zheng Yuan, Yijia Liu, Chuanqi Tan, Songfang Huang, and Fei Huang. 2021. Improving biomedical pretrained language models with knowledge. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 180–190.

Zheng Yuan, Zhengyun Zhao, and Sheng Yu. 2020. Coder: Knowledge infused cross-lingual medical term embedding for term normalization. *arXiv preprint arXiv:2011.02947*.